**The Use and Abuse of Formal Models in Political Philosophy**

**A Tendentious Tutorial in**

**Five Parts**

**by**

**Robert Paul Wolff**

**Introductory Remarks**

The purpose of this little book is to introduce you to the technical foundations of a number of formal methods of analysis that have come to play a large role in the writings of philosophers, economists, political theorists, legal theorists, and others, and then to show you how these formal methods are **misused** by many of those theorists, with results that are conceptually confusing and quite often ideologically suspect. I am going to expound these materials carefully and with sufficient detail to allow you to master them and make your own judgments about the appropriateness of their use.

There are three distinct bodies of material with which we shall be dealing. Each has grown out of a different intellectual tradition, uses different methods of formal analysis, and finds different application by philosophers, political, theorists, and so forth. Quite often they are confused with one another, and my impression is that the people who use them frequently do not understand the distinctions among them, but we shall treat them separately.

The first body of material is Rational Choice Theory. When people talk about **maximizing utility** or calculating the **expected value** of an alternative or **discounting an outcome by its risk**, they are drawing on Rational Choice Theory.

The second body of material is Collective Choice Theory. When people talk about **the paradox of majority rule** or **Arrow's Theorem** or **Pareto Optimality** they are drawing on Collective Choice Theory.

The third body of material is Game Theory. When people talk about **strategies** or **zero sum** or **prisoner's dilemma**, they are drawing on Game Theory.

I am going to ask you to be patient, because this is going to take a while. By the time we are done, this tutorial will have become a short book. As we proceed, I shall make some reading suggestions for those of you who wish to pursue the subject in greater depth, but everything you will need to know to follow my exposition will be contained in these pages.

The order of exposition is going to be as follows:

1. Some preliminary technical matters, principally concerning different kinds of orderings.

2. The elements of Rational Choice Theory.

3. The elements of Game Theory, including the von Neumann and Morgenstern formal development of the axioms of carcianl utility functions, and a statement and explanation [but not formal proof] of the fundamental theorem of Game Theory.

4. The elements of Collective Choice Theory, including a formal proof of Kenneth Arrow's General Possibility Theorem, which is the central formal result in the field, followed by

5. Applications, including a skeptical look at the so-called Prisoner's Dilemma, and a formal analysis of John Rawls' claims concerning choice in the Original Position in *A Theory of Justice*.

**AN APOLOGY: I cannot see your eyes, so I cannot tell when they glaze over, either from boredom because I am going too slowly, or from confusion because I am going too fast. I shall do my best to choose a speed and technicality of exposition that strikes a balance between these two extremes, but I apologize now if I fail.**

**Part I. Preliminary Technical Matters**

**A. Scales of Measurement** [For more detail, see S. S. Stevens, *Handbook of Experimental Psychology*, originally published in 1951 but re-issued and updated.]

Let us suppose that we have a finite set of discrete elements of any sort, which we will call S =(a, b, c, ..., n). The elements might be different amounts of money, different flavors of ice cream, different bowls of ice cream [not the same thing, of course], different candidates in an election, and so forth.

We may wish to impose an ordering on the set. The very simplest ordering we can impose is a **nominal ordering**, or a **labeling**. To each element, we assign a label or name [hence "nominal"]. Two or more elements may receive the same name, but each element receives only one name. Such an ordering is said to be **complete** if every element in S is labeled. The ordering creates what are called **equivalence classes**, which is to say, subsets of elements all of which bear the same label or name. This labeling exhaustively and mutually exclusively divides S into subsets. Obviously, two elements are in the same equivalence class if and only if they have the same name. Every element is in one, and only one, equivalence class. With a **nominal ordering**, nothing more can be deduced from the labeling than the simple fact that two elements are in the same equivalence class if and only if they bear the same label. The essential fact about this very simple measure is that it is **complete**. Every element bears a label. For any two elements, either they are in the same equivalence class or they are not. Trivially, each element is in the same equivalence class with itself. Thus, every element is in some equivalence class.

The next step is to introduce a binary relation, R, over the set of elements. xRy is construed variously as meaning "x is equal to or greater than y," or "(someone is) indifferent between x and y or prefers x to y," or even "x is hotter than or is the same temperature as y," and so forth. All of these have the same formal structure. Let us suppose the following two propositions are true for R and for S:

(i) for all x and y in S, xRy or yRx. This says that R is **complete**. Notice that from this, it follows that for all x, xRx. [Just as a trivial exercise, here is how we prove that xRx. Since for any x and y, xRy or yRx, take the case in which x=y. Then substituting, we have xRx or xRx, which is logically equivalent to xRx. That is the sort of baby logic steps we will be taking many of in what follows]. This property of an element bearing a relation to itself is called **reflexivity**, and a relation of which it holds is said to be **reflexive**.

(ii) for all x, y, and z in S, if xRy and yRz then xRz. This property is called **transitivity,** and it will turn out to be the single most important property of relations like R.

Just to be absolutely clear what we are talking about here, suppose we interpret the relation xRy to mean (someone) prefers x to y or is indifferent between x and y. Then (i) says that for any two members of the set S, the person in question either prefers x to y or is indifferent between them, or else prefers y to x or is indifferent between them. If this is still a bit puzzling, think of x and y as real numbers and R as meaning "is equal to or greater than." (ii) says that if the person in question prefers x to y or is indifferent between them, and also prefers y to z or is indifferent between them, then that person also prefers x to z or is indifferent between them.

A binary relations like R is said to establish a **weak ordering** on the set S. It is weak because it allows for indifference. Starting with the relation R, we can also define a relation P on S, like this: xPy means xRy and not yRx. P here stands for "prefers," and a relation like P is said to establish a **strong ordering** on the set S. To get an intuitive handle on these very important little symbols, think of it this way. xRy says that x is at least as good as y, and maybe better. xPy says that x really is better than y [whatever "better" means here.] So R is weak and P is strong. Later on, when we come to Collective Choice Theory, we will be saying a lot about weak and strong orderings.

A relation, R, over a set, S, for which i) and (ii) hold is said to be an **ordinal ordering**. In discussions of these matters in philosophy, economics, and political theory, it is often taken as a fundamental test of a person's rationality that his or her preferences exhibit at least an **ordinal ordering** over all available alternatives. Some economists, using what is called a theory of "revealed preference," even argue that everyone **must** have a preference structure that at least satisfies the first condition, and thus is **complete**, because confronted with any two alternatives, x and y, a person will either choose one, thus showing that she prefers it to the other, or else will be indifferent between the two. But that, I suggest, is a covertly tendentious thesis made more plausible by the formalism.

By the way, "**ordinal"**  because the ordering merely establishes which of the elements is *first, second, third, fourth*, etc. according to the relation R, and these are what are called "ordinal numbers."

It may not be obvious at first glance, but preference structures do not always exhibit transitivity, and hence are not even ordinal. Indeed, the casual assumption of transitivity is actually an enormously powerful and simplifying assumption. Let me give an elementary and non-controversial example here, and save the controversial examples for later. All of us, I assume, have had our eyesight checked at the optometrist's office. You shut one eye, the room is darkened, and you look through a complicated gadget at a chart of rows of letters, each line smaller than the one above. The doctor flips lenses in front of your open eye, and asks "Which is clearer, one, or two?" Sometimes you can see a difference, and sometimes you just say, "They are the same." The two lenses may actually have different degrees of magnification, but the difference is simply too small for you to notice. Experimental psychologists say that the difference between the two is then below your "minimal discriminable difference," or MMD. Now, it is obvious that with a little work, the optometrist could line up a series of lenses, each successive pair of which falls below your MMD, but the first and last of which are clearly discriminable. If we interpret R in this case to mean "is clearer than or is equally as clear as," it would be true that for any adjacent pair, m and n, mRn and nRm, but for the first, a, and the last, q, it would not be the case that aRq and qRa. In other words, the relation "is clearer than or equally as clear as" would not be transitive. The same thing might manifestly be true of someone's preferences. What all this means is that it is very powerful and quite probably false to assume that someone has a transitive preference ordering over a set of available alternatives. But people who use this sort of formalism almost never realize that fact. Indeed, it is quite often the case that people introduce this formalism without even feeling any need to say that they are assuming transitivity. This is a simple example of what I mean when I say that the formalism can conceal powerful and dubious assumptions.

Ordinal preference orders encode the order of someone's preferences, but not the intensity of that preference. Compare voter A with voter B in the 1992 presidential election. Voter A is a fanatic George H. W. Bush supporter. She doesn't really like either Clinton or Perot, but despite Perot's kookiness, prefers him by a hair to Clinton. Voter B is torn between Bush and Perot, neither of whom he loves, but he finally decides to go with Bush. He hates Clinton and wouldn't vote for him even if Mao Tse-Tung were the alternative. These voters have identical ordinal preference structures: Bush first, Perot second, Clinton third. That is all you need to know to figure out how they will vote, but obviously for all sorts of other purposes this ordinal preference ordering fails to embody a great deal of important information. In particular, this ordering will not tell you how either voter might behave in other political contexts besides voting, such as donating money, working for a campaign, lobbying, and so forth.

This is as good a place as any to call into question the easy assumption that the possession of a complete ordinal preference structure is the most elementary test of one's rationality . A great deal is at stake here, much more than you might think. Let us start slow. The theory of rational choice has its roots in analyses of gambling behavior, of economic behavior, and -- to some degree -- of political behavior. Now, when we are talking about the way in which professional gamblers decide how to play their cards or place their bets, it makes sense to assume that they can define a complete preference order over the available alternatives. That is to say, the various possible outcomes offered by a gambling game are plausibly described as commensurable with one another. The outcomes are, after all, simply different wins or losses of amounts of money. The same is true of people engaged in economic activities. But these are relatively limited and specialized arenas of human activity. There are many other arenas in which it is not so obvious that rational individuals have complete preference orders over available alternatives.

Consider, as an example, the terrible choice presented to the central character in William Styron's novel *Sophie's Choice*. [I know the story from the movie of the same name, starring Meryl Streep.] A Gestapo gauleiter overseeing the loading of Jews onto trains taking them to the death camps offers Sophie a choice. She may save one of her two children from certain death, but she must choose which one will survive. His posing of this choice is clearly an act of satanic sadism. There are two ways of thinking about this situation. The natural, and I suggest, rational way to think about it is as a tragedy in which a woman is presented with a terrible situation that will destroy her life no matter what she does. To choose either child is impossible. To fail to choose one is to condemn them both to death. Religion may have something useful to say about this situation. Literature may. Perhaps nothing can. But for sure Rational Choice Theory is no help. But Rational Choice Theory says that she must have some preference order or other over the three outcomes, and her choice, whatever it is, reveals that preference.

Let me put this in a summary fashion, and ask you to think about what I say. Perhaps later on we can discuss it. The assumption of a complete ordinal preference order is presented in the literature as an innocuous premise that gets the more complex and interesting arguments going. But in fact it is, covertly, a highly questionable proposal to extend a form of economic rationality into areas in which it arguably does not belong. By accepting the formalism, someone unwittingly buys into this powerful encroachment of the economic into arenas of human experience in which it has no place. Imagine coercing a man into acting dishonorably, and then saying that his agreement reveals exactly what price he places on his honor. It would be more true to the human reality to say that by this act of coercion, you have besmirched his honor, which henceforth is worth nothing to him. The outcome of the choice you forced on him is not a rational choice but shame.

The defining characteristic of capitalism is the reduction of all human activity to market relations. Too often, Rational Choice Theory functions as a covert and seductive rationalization of the capitalist ethos, which then seems, because of the apparent neutrality of the formalism, to be equivalent to rationality *tout court*.

It is not necessary to limit ourselves to complete orderings -- orderings which establish the individual's or society's preference for any two alternatives whatever. We can also define **partial orderings,** and these have in fact played an important role in Economics and other disciplines. I will only say a few words here, and return to this subject down the line. The *Sophie's Choice* example has shown us that sometimes individuals cannot say, for two alternatives, which one they prefer. It is not that they are indifferent between the two. The two are simply, in their minds and hearts, not comparable. How many lives is it worth to save the only score of Bach's B Minor Mass? The question makes no sense to us, no matter what phony scenarios we cook up in a philosophy essay.

A similar problem arises when we are trying to compare different social distributions of wealth. If Situation B offers everyone more wealth than Situation A, then we can be pretty sure there will be unanimous agreement that B is better than A. Indeed, if people are willing not to be envious of what others get, then we might be able to secure unanimity for the proposition that B is better than A if B offers everyone at least as much as A does, and offers at least one person more. [Why begrudge her the extra if it isn't coming out of your share?] But what about the case in which B makes some people better off and others less well off than they were in A? There may just be no answer in that case.

Thanks to Vildredo Pareto [1848 - 1923], when B makes everyone better off than they were in A, we say that B is Pareto Preferred to A. Obviously, if B is Pareto Preferred to A and C is Pareto Preferred to B, then we should expect that C will be Pareto Preferred to A. So this Pareto or Unanimity ordering is transitive but not complete. If some way of distributing things is such that there is no alternative distribution that is Pareto Preferred to it, then we say that it is **Pareto Optimal**. Don't be misled by the enticing sound of the word "optimal." If we assume that everyone has **positive marginal utility** for money, so that taking even a little bit away from someone makes her less well off, then a social distribution that gives everything to one man and nothing at all to anyone else is Pareto Optimal, because any re-distribution will involve making at least one person worse off, namely the person who had everything and now has slightly less. In case you are wondering why this matters, I will just point out that when economists describe a market as **efficient**, they mean that it produces a Pareto Optimal outcome. Not too heart warming.

So much for **ordinal preference orders,** at least for the moment. Now things get somewhat more complicated, but also a good deal more important. The next step up, after **nominal** and **ordinal** orderings, is **cardinal** orderings. Since this is going to require a little technical work, let me first explain what is at stake. Both Rational Choice Theory and Game Theory [but not Collective Choice Theory] involve talking about people doing something called "maximizing expected utility," or "discounting the value of an outcome by its risk" and so forth. These calculations require that we be able to assign cardinal numbers, or magnitudes, to different outcomes or alternatives, and that we be able then to do things like adding them, subtracting them, multiplying them by other numbers, etc. Now, you cannot add or subtract or multiply or divide ordinal numbers. It makes no sense to ask, "Is Second the average of First and Third?" in the way that you might ask "Is 2 the average of 1 and 3?" If you have ever been involved in trying to work out a system to decide which team in a track meet has won over all, or which country has won over all in the Olympics, you will understand this. Does a whole raft of silver and bronze medals count for more or for less than a small pile of gold medals? Are a gold and a bronze equal to two silvers? The questions are meaningless. To carry out any of these calculations, you need an **interval scale**, also called a **cardinal ordering**.

An **interval scale** is an assignment of numbers to the elements of an ordinal ordering in such a way that the **intervals** are equal [hence "interval scale." This is actually a gross simplification of the correct definition, but I don't want to scare people away, and this will suffice.] A good example is the Fahrenheit temperature scale. The elements here are, let us suppose, readings provided by a thermometer of the temperature of different bodies of water. We can first impose a **nominal ordering** by grouping together the bodies of water that are [or maybe feel] the same temperature. We then impose an **ordinal ordering** by arranging the equivalence groups in a hierarchy from hottest to coolest. Thus far, all we have is the information that this body of water is hotter than that one, or maybe that this body of water is at least as hot as that one [i.e., weak rather than strong ordering]. Now, suppose we can actually answer the following question for any four bodies of water, a, b, c, and d: **Is the difference between the temperature of a and the temperature of b at least as large as the difference between the temperature of c and the temperature of d?** Notice I said any four bodies of water. In other words, I am asking about **intervals** of temperature, not just temperatures. If I have enough information to answer that question for any tetrad of bodies of water, then I can define a **cardinal** measure of temperature,. I can say, for example, using the Fahrenheit scale, that the **difference** or **interval** between fifty degrees and sixty degrees is the same as the **difference** or **interval** between twenty degrees and thirty degrees. So it makes sense to say, "It is ten degrees cooler today," regardless of what the temperature was yesterday.

We are here performing an arithmetic operation on the labels assigned to the elements of the set [namely subtraction]. But you cannot perform arithmetic operations on ordinal numbers. For that you need cardinal numbers [i.e., real numbers] like 1, 2, 3, and 4. So, this sort of scale is called a **cardinal** scale.

The last step, which is only important for a few purposes, is to define what is called a **ratio ordering** or a **ratio scale**. A ratio scale is just like a cardinal scale but with one thing added: with a ratio scale, we have enough information to say, for any four elements of our set, a, b, c, and d, whether the ratio of a to b is equal to or greater than the ratio of c to d. Or, going back to the symbolism we used above, whether a/bRc/d. Now a little experimentation will show you that a **ratio** scale requires that you be able to identify some point as the zero point, or origin. A Fahrenheit temperature scale is **not** a ratio scale. The zero point in the Fahrenheit scale is chosen arbitrarily. Therefore, it makes no sense at all to say that in a Fahrenheit scale, the ratio of twenty degrees to ten degrees is the same as the ratio of eighty degrees to forty degrees. [For those of you who know some Physics, that sort of statement does make sense in a Kelvin scale of temperature, where the zero point is what is called absolute zero.]

Why am I going on about this? Well, for one reason, because it will turn out, way down the line, that without knowing this stuff you cannot understand what a zero-sum game is!

**B. Transformations**

Now we are going to talk about transformations. Technically, a transformation is a one-one mapping of a set onto another set, but we can just think about a transformation as a rule for assigning new labels or numbers to a set of elements.

1. **A permutation** is a re-assignment of the labels attached to the elements that preserves their grouping into equivalence classes. Initially, you will recall, we attached labels to the elements of our set, S. Two elements that got the same label were then in the same equivalence class. So if we label people by their last names, all the Millers go together, all the Taylors go together, and so forth. We could now relabel everyone, say by translating their names into another language [so that all the Millers become Muellers, and all the Taylors become Schneiders.] That would change everyone's name, but it would not change the groupings into equivalence classes. All the Millers were together before, and they are still together now that they are all Muellers. What is more, no two people who were in different groups before are in the same group now. The official jargon for this state of affairs is that **the labeling is invariant under a permutation**.

2. **A monotone transformation** is a re-labeling that preserves an ordinal ordering. Suppose we take the items labeled first, second, third, and fourth, and now label them fifth, eleventh, nineteenth, and fortieth. No information in the original ordering has been lost, and none has been gained. In the formalism of the relation R, for any two elements a and b in S that have been relabeled a' and b' respectively, aRb if and only if a'Rb'. So the ranking has not been changed by the transformation. Again, the official way to say this is that **the ordinal ordering, R, is invariant under a monotone transformation**.

3. **An affine transformation** is a transformation that preserves a cardinal ordering. A **linear transformation** of a relation R on a set of elements S = (a,b,c,....n) is a relabeling of each element a in S such that the new label, a' equals the old label a times some constant plus another constant. Or: a' = aq + r. This used to be called a **linear transformation** because the expression (a' = aq + r ) is the formula for a straight line drawn on the a and a' axes. A little elementary algebra will show that this transformation preserves an interval scale or cardinal ordering on the elements of S. Remember: This means that it preserves equality of intervals between pairs of elements. Here is how we prove this:

Take four elements, a, b, c, and d, such that (a-b) = (c-d), Now impose a **linear transformation** on S. That means:

a' = aq + r

b' = bq + r

c' = cq + r

d' = dq + r.

Notice that we have imposed the same linear transformation on each element. In other words, the constants q and r are the same in each case.

By hypothesis (a - b) = (c - d)

Substituting the transformed labels, we get (aq+r - bq-r) = (cq+r - dq-r)

or: (aq-bq) = (cq-dq)

Dividing both sides by q, we get (a-b) = (c-d) Ta Da!

So long as you re-label the elements of S so that the labels satisfy the equation a' = qa+r, for any a in S, it makes no difference which set of labels you use, because they all encode the same information. This is what it means to say that a **cardinal ordering is invariant under an affine transformation**.

An affine transformation does two things. It changes the size of the intervals [but not the equality of different pairs of intervals], and it changes the zero point. The classic example of an affine transformation is the formula for converting temperature from Fahrenheit to Centigrade. The formula, as everyone knows who travels to Europe and wonders whether to wear a sweater or not, is F degrees = 9/5 Centigrade degrees + 32. So, if *Le Monde* says it is going to be 20 degrees today in Paris, that means 9/5(20) + 32 or 68, so no sweater. Zero degrees in centigrade is the temperature at which water freezes, but in Fahrenheit, that is 32 degrees. And so forth. Each degree Centigrade is equal to 9/5 Fahrenheit degrees. Why does this matter? Once again, it will turn out to be crucial when we come to give a correct definition of a zero sum game, and for many other purposes besides.

Finally, and uninterestingly, a **ratio transformation** is a transformation of the form a' = qa. This the formula for a line that goes through the origin of the graph. The transformation just changes the size of the intervals but does not change the zero point. And obviously, a'/b' = qa/qb = a/b. This one doesn't matter much, but I put it in for completeness' sake.

O.K. We have **nominal, ordinal, interval, and ratio scales**, and we have **permutations, monotone transformations, affine transformations, and ratio transformations**.

Now, as Portnoy's analyst says in the last line of the novel, let us begin.